FISS: Function identification of subway stations based on semantics mining and functional clustering

Article *in* IET Intelligent Transport Systems · January 2018
DOI: 10.1049/iet-its.2017.0316

7 authors, including:

Jinzhong Wang
Dalian University of Technology
15 PUBLICATIONS **63** CITATIONS
SEE PROFILE

Xiangjie Kong
Dalian University of Technology
73 PUBLICATIONS **462** CITATIONS
SEE PROFILE

Azizur Rahim
Dalian University of Technology
26 PUBLICATIONS **242** CITATIONS
SEE PROFILE

IET Intelligent Transport Systems

Research Article

IET Journals
The Institution of Engineering and Technology

FISS: function identification of subway stations based on semantics mining and functional clustering

ISSN 1751-956X
Received on 1st October 2017
Revised 12th December 2017
Accepted on 6th January 2018
doi: 10.1049/iet-its.2017.0316
www.ietdl.org

Tao Tang[1], Xin Dong[1], Jinzhong Wang[2,3], Xiangjie Kong[2] ✉, Azizur Rahim[2], Xuannian Yu[2], Yulin Li[2]

[1]*Chengdu College, University of Electronic Science and Technology of China, Chengdu, People's Republic of China*
[2]*School of Software, Dalian University of Technology, Dalian, People's Republic of China*
[3]*School of Management and Journalism, Shenyang Sport University, Shenyang, People's Republic of China*
✉ *E-mail: xjkong@ieee.org*

Abstract: In modern cities, the subway system plays an important role in carrying a large proportion of passenger transport. However, there still remain some issues on how to accurately identify the regional functions of subway stations. In this study, the authors propose an approach named FISS for identifying the functions of subway station regions based on semantics mining and functional clustering. First, they extract the passenger's travel patterns of each subway station based on the smart card transaction data and Shanghai subway network data and calculate the relative point of interest (POI) contents of each subway station by using Shanghai POI data, then feed the two above-mentioned results into latent Dirichlet allocation model for obtaining the mobile semantics and the location semantics separately. Furthermore, they carry out standardisation after combining the two semantics, then extract the functional characteristic vectors of subway stations by conducting sparse principal components analysis, and cluster these vectors by using the improved *k*-means algorithm. At last, they visualise the result after taking subway station's function identification by the interclass passenger flow transfer, the distribution of geographical function proportion and the similarity of inter-cluster. The results demonstrate the accuracy and efficiency of the proposed approach compared with other existing methods.

1 Introduction

As the rapid development of information technology, the construction of smart city is more closely connected with the deployment of big data [1]. It has brought a great convenience to urban inhabitants in their daily travel with the extensive use of sensor technologies, the intelligent transport systems and the location-based information technology services, thus a large amount of city data has produced every day that we can collect and take a good use of the information data from the human movement trajectory, social activity and so on. The research of regional function identification aims to mark the main function of each region in a city by extracting the commuting patterns from the datasets which contain the relevant information such as the trajectory of public vehicles, the records of an intelligent transport system. It can show the valuable reference information for the administrators and planners of the city for optimising the city resource allocation to lessen the problems like traffic congestion, environmental pollution and the division of all construction lands.

The massive urban datasets contain the concealed valuable information which can be used for population mobility detection [2], research on multi-traffic behaviour [3–5], identification of different urban functional regions [6] and so on. In urban computing, for obtaining these hidden information, we need to implement specialised data mining methods in the data from the complex structure and different sources. Data mining is a cross-discipline in computer science which combines statistics, artificial intelligence, machine learning and the processing of finding the patterns from datasets in a database system. The function of data mining is to extract the main information from datasets and transform the information into intelligible structures for the future work.

The subway has become the optimal choice in modern city transportation system by virtue of its vast capacity, convenience, on time, low environmental pollution and so on. On one hand, each of the region in subway stations is regarded as a central landmark of the city that the subway system promotes the communication among the central regions. On the other hand, the subway has accelerated the development of the regions around the subway line and some new functional areas have been clustered under this situation. As everyone knows, different urban areas will be evolved into their corresponding functions with the development of a city, which can fit the needs of the residents. The urban areas are probably reshuffled by the planner or naturally formed by the human lifestyle and it can be changed during the process of urban development. The regions around subway station are the typical representative of the above processes.

In this paper, we use smart card transaction data in Shanghai subway, Shanghai subway basic data and Shanghai points of interest (POIs) data, and utilise semantic analysis algorithm and functional clustering algorithm for identifying the regional functions of subway stations based on data mining. We can know the distribution of the core functions, the development process of the city by digging the functions of subway stations, and our work provides a valuable reference for urban transportation planning, regional development planning and urban resource allocation, all of which are meaningful to construct the smart city.

Our major contributions can be categorised as follows:

i. We propose an approach that can identify the major function of subway stations in the city using traffic card transaction data and POIs data.
ii. We conduct the station semantic and mobile semantic discovery to obtain their potential topic distribution vectors, respectively, by using latent Dirichlet allocation (LDA) model algorithm.
iii. The generated mobile semantic and station semantic matrices are spliced and *Z*-score normalised, and then the sparse principal component analysis (SPCA) is used to process the resulting matrix to obtain the station-function matrix.
iv. We use *k*-means clustering algorithm to get the functional station clustering results and conduct map visualisation for these clusters and identify the function of subway stations through the results of intercourse passenger flow transfer, geographical function proportion distribution and cluster similarity.

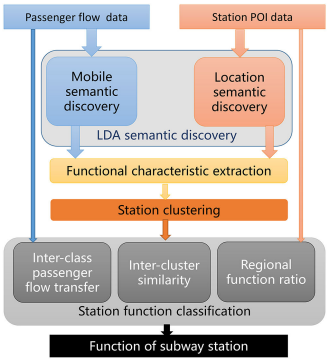


Fig. 1 *Framework of subway station functional identification*

We obtain the results of subway station classification by implementing this approach with processing the smart card transaction dataset in Shanghai subway and the Shanghai POIs dataset, and in the contrast experiment, we conduct the experiment with changing the inner algorithms into other alternative algorithms in our approach, thus we use term frequency-inverse document frequency (TF-IDF) to process Shanghai POIs dataset, choose the affinity propagation (AP) algorithm and MeanShift algorithm for subway stations clustering. In addition, we also repeat the experiment with decreasing one of its steps by only mining the single mobile semantics or location semantics, conducting the experiment without implementing SPCA. After that, we compare the clustering results of these changed approaches with the clustering results of our approach.

The remaining part of this paper is organised as follows. The related work of regional function identification and urban computing based on passenger flow are introduced in Section 2. Section 3 elaborates on the framework of FISS we proposed for identifying the regional function of subway stations. Section 4 shows the subway station clustering results of FISS and the analysis of these results. Section 5 verifies the validity of FISS by comparing the result of FISS with the result of a changed procedure under the same framework from several aspects (such as changing the clustering algorithm). At last, this paper is concluded and future works are highlighted in Section 6.

2 Related work

We introduce the relevant works about regional function identification and urban computing based on subway passenger flows in this section.

2.1 Regional function identification

In recent years, exploring the distribution of functional regions plays an important part for realising smart city. Karlsson *et al.* [7] proposed an overview of economic clusters and clustering idea. They defined the characteristic of function regions as a cluster which includes the beneficial economic activities. In traditional functional region identification, Unsalan [8] explored the development of urban regions by using high-resolution satellite shooting images. However, this method cannot meet the requirement about timeliness and low cost in urban computing. Zhi *et al.* [9] proposed a model based on low-rank estimation and thus obtained five typical functional clusters. In their work, they defined a series of patterns of latent space–time activities, provided a new inspiration for future research.

Recently, Assem *et al.* [10] proposed a new approach for functional region modelling which has taken into account the time changes of the regional label, and also provided an approach for comparing the effect of three clustering methods including hierarchical clustering, *k*-means clustering and spectral clustering. Kraft and Marada [11] introduced the conception of the local minimum value of transportation intensity based on car traffic. Fan *et al.* [12] used an unsupervised feature learning algorithm to identify the usage scenario of land based on remote sensing data. Furthermore, Yin *et al.* [13] proposed an approach for separating the regional boundary based on the human interaction data which included more than 69 million Tweet location information. Rudianac *et al.* [14] used the convolutional neural network to identify the functional regions based on social multimedia data and explored the distribution of potential themes from their identification.

There remains some research on human mobility for reflecting the function of urban regions. Qi *et al.* [15] identified regional functions by analysing the change of get-on/off amount from taxi trace data (GPS data) and classified them into three types. Yuan *et al.* [6] introduced LDA model algorithm into semantic mining for identifying different regions. Sarkar *et al.* [16] created a functional matrix for revealing the links between the potential functions of the regions.

2.2 Urban computing based on passenger flow

In recent years, urban computing has become a research hotspot for processing the big traffic data more efficiently. Therewith, the relative research of urban computing on passenger flow is a vital part in this field. Bhaskar *et al.* [17] proposed a method to segment transit passengers individually using smart card data and mined the single card users travel pattern using density-based spatial clustering of application with density-based spatial clustering of applications with noise algorithm. Smart card users are segmented into four typical classes by the a priori market segmentation approach which includes regular origin–destination passengers, time regular passengers, commuting passengers and irregular passengers.

Zhao *et al.* [18] proposed a novel approach for discovering the patterns of passenger's route selection by using the data of subway automatic fare collection system, they displayed a method for calculating the probability of the origin and destination point with multiple routes chosen in the complex subway network, and they obtained the whole possible and efficient planned routes of a single passenger by matching each smart card and using records of the passengers and the subway operation log.

Zhang *et al.* [19] proposed a new approach for extracting the spatiotemporal streams of passenger flow from subway passengers' inbound and outbound data, which can contribute to traffic prediction, planning, scheduling and so on.

Itoh *et al.* [20] designed a novel visual fusion analysis system for representing the changes in passengers' behaviour and the unconventional station by using smart card data of Tokyo subway and Tweet data. Ni *et al.* [21] developed a hashtag-based event detection algorithm for detecting various events based on the hashtags of social media (Tweet) data. Furthermore, they proposed a parametric and convex optimisation-based approach named optimisation and prediction with hybrid loss function which fused the linear regression and the prediction results of seasonal autoregressive integrated moving average model, for forecasting passenger flow of events nearby a subway station under event conditions by using social media data.

From the related work above, we find that the precision of regional function identification is not good enough. There exists little research on regional function identification of subway stations, we are the first to utilise the hybrid semantic model to identify the regional function of subway stations.

3 Framework of FISS

We propose an approach named FISS in this section for identifying the function of subway stations. Fig. 1 shows the framework of FISS.

The whole process of FISS is divided into four modules:

i. Semantic mining based on LDA algorithm. Taking the initial passenger flow travel data and POI data of station area as input, we use the station-mobile semantic mining, station-location semantic mining and document-subject semantic mining. We also take the distribution matrix of passenger travel pattern and POI relative content matrix as the input of LDA, and implement the LDA topic model for mining the static semantics of stations.

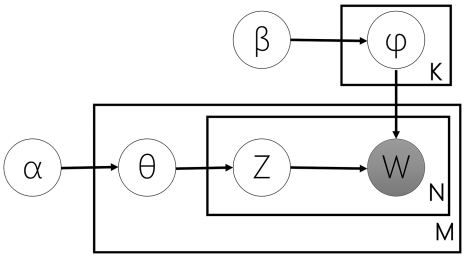


Fig. 2 *Probability graph model of LDA, α is the parameter of the Dirichlet prior on the per-document topic distributions, β is the parameter of the Dirichlet prior on the per-topic word distributions, θ_m is the topic distribution for document m, φ_k is the word distribution for topic k, z_{mn} is the topic for the nth word in document m and ω_{mn} is the specific word*

ii. Functional characteristics extraction of stations. We combine mobile semantics and location semantics, and extract the functional feature vectors of each station by using zero-mean (Z-score) normalisation and SPCA.
iii. Station clustering. For identifying the function of each station, we need to categorise stations with similar functionality, which needs clustering the functional vectors of stations. We choose the optimal *k*-means algorithm for clustering which has the optimal performance.
iv. Station function identification. Identifying the function of clustering results from three parts including inter-class passenger flow transfer, inter-cluster similarity and regional function ratio. Thus, we obtain the final result after marking the function of stations.

In the following sub-sections, the four modules will be described in detail, respectively.

3.1 Semantic mining based on LDA

Regional function mining is an important subject in urban computing, functional identification of subway station is a new perspective we proposed in this direction. We try to reveal the location semantics of the station function, which includes residential areas, commercial areas, education areas and so on. It should be noticed that the final result will assign a specified category to each station, but it does not mean the station can only provide the specified function. In reality, regional functions in a city are heterogeneous. Therefore, we categorise the most significant core function after comparing with other functions.

From the angle of transport function of the subway system, subway system carries a big part passenger flows of the city, it also becomes a major choice for citizens. The passenger flow data of each subway stations with different functions can show its own characteristics, that is the passenger flow data can reveal the function of a station, we named this as mobile semantics. From the angle of urban geographical function, the subway stations are set up in the core regions of a city that these regions can be linked by subway lines. Conversely, the subway lines also can promote the development of the regions where the stations located, and this feature highlights the static function (functions of buildings, and other static places) around the subway stations, we name that as location semantics. The above two datasets reflect the subway station functions from two angles, and this section mainly introduces how to reveal this two semantics.

3.1.1 Mobile semantics mining: For obtaining the mobile semantics from the passenger flow data, we conduct mobile semantic mining by implementing LDA subject model. We build the passenger travel pattern distribution matrix based on the passenger flow data and take this matrix as the input of LDA model, then we obtain the mobile semantics from the LDA model.

LDA is a generative probabilistic model for a corpus [22]. LDA is the modification of probabilistic latent semantic analysis (PLSA) added the prior distribution [23]. It considers that each article contains a number of potential topics and occupies a different proportion. Each word in the document is generated by a topic. The probability graph model of LDA [24] is shown in Fig. 2. The outer box represents documents, while the inner box represents the repeated choice of topics and words within a document. M denotes the quantity of documents, N denotes the quantity of words in a document.

LDA assumes the generative process of a document in the document set D with M documents each length N_i is in the following:

i. Choose $\theta_i \sim \mathrm{Dir}(\alpha)$, of which $i \in \{1, \ldots, M\}$ and $\mathrm{Dir}(\alpha)$ is a Dirichlet distribution with a symmetric parameter α which typically is sparse ($\alpha < 1$).
ii. Choose $\varphi_k \sim \mathrm{Dir}(\beta)$, of which $k \in \{1, \ldots, M\}$ and β typically is sparse.
iii. For each of the word positions i, j, where $j \in \{1, \ldots, N_i\}$, and $i \in \{1, \ldots, M\}$.

(a) Choose a topic $z_{i,j} \sim \mathrm{Multinomial}(\theta_i)$.
(b) Choose a word $\omega_{i,j} \sim \mathrm{Multinomial}(\varphi_{z_{i,j}})$.

In the application of LDA, it takes the distribution of a word in each document of the observed document sets as the input of LDA, and it can work out the distribution of each document's topics from LDA model.

In our model, passenger flow data is a collection of travel records. Each record J consists of the following five items: start station S_L, destination station S_A, start time T_L, arrive time T_A and date D, thus, $J = (S_L, S_A, T_L, T_A, D)$. These travel records reflect the kinds of human social commercial activities, for instance, on weekdays, most of passenger flow of the happened activities from residential regions to commercial regions during the morning rush-hour crush are on duty, and at weekends, passenger flows in the morning from residential regions to commercial regions are probably on shopping or in leisure. Taking through these activities from the angle of station's passenger flows, stations of the living region are showing departure peak in the morning of weekdays, and for the commercial regions, the stations are showing arrival peak. Meanwhile, there are many other functional regions like education regions, development regions and so on. Each specific functional region shows its special characteristic on departure flow and arrival flow during the period of one day, of which these characteristics can reflect the function of the station as a functional 'fingerprint'.

By modelling the analysis process, we can define the passenger flows which can reflect its latent function as the travel pattern P from the angle of station. The frequency of a specifical travel pattern P can reflect the function of the station. The next step is to determine what information to be included in the P.

From the above sentence 'stations of the living region are showing departure peak in the morning of weekdays', we can find that a travel pattern P that appears in a station includes the following information: whether the data D of a travel record J is happened on weekdays D_W or on weekends D_H. Whether the travel record is departed L or arrived A. Whether the period t of the day is happened at the period $t_{(T_L)}$ or the period $t_{(T_A)}$ (In departure patterns, $t_{(T_L)}$ represents the period of departure T_L. In arrival patterns, $t_{(T_A)}$ represents the period of arrival T_A.), specially, we divide the day by 1 h. Whether is the station from which the station S_A arrives or is from the departure station to the station S_L.

Thus, there are two patterns P that can be extracted from one travel record J. Assumed $J \cdot S_L$ as station S_i no. i, $J \cdot S_A$ as station S_J no. j. Then, for station S_i, pattern P includes $P(D_W | D_H, L, t(T_L), S_A)$, and for station S_j, pattern P includes $P(D_W | D_H, A, t(T_A), S_L)$.

The analysis above is the process of extracting pattern P from one travel record J, and we need to obtain the travel pattern's frequency of each station. Then, we pass the frequency of all subway stations through a $m \times n$ station-travel pattern matrix $\boldsymbol{M}_{SP}$, of which m is the total number of stations, n is the total number of all possible travel patterns. In matrix $\boldsymbol{M}_{SP}$, the element $M_{SP} \cdot m_{ij}$ is

the number of times that the pattern P_j in the station S_i appears, of which $i = 1, 2, 3, \ldots, m$, $j = 1, 2, 3, \ldots, n$.

We find the process of the mobile semantics of subway station is completely in conformity with the distribution which calculated by LDA model by analogy. In our model, both α and β are input parameters in LDA, where α corresponds to the station-function matrix, β corresponds to the function-travel pattern matrix, θ_m corresponds to the function distribution of station m and φ_k corresponds to the travel pattern distribution for function k, z_{mn} corresponds to the function for the nth travel pattern in station m and ω_{mn} corresponds to the specific travel pattern.

The analysis above has clearly clarified that the distribution of the travel pattern on each station can reflect the function of the station, and the same as the distribution of the word in each document can reflect the latent topics of the document. It shows the process of this analogy in Table 1.

After building matrix $\boldsymbol{M}_{SP}$, we conduct semantic mining on the passenger flow data for obtaining its mobile semantics by implementing LDA model. Therefore, we take the station-travel pattern matrix $\boldsymbol{M}_{SP}$ as the input of LDA, then we can obtain a $m \times k$ station-function matrix, of which m is the number of subway stations, k is the number of the latent functions. Each row of this matrix represents the distribution of k latent mobile semantics in one station. It is worth mentioning that the obtained latent semantics are not the specific functions that can be explained eventually, so the value of k does not represent the number of final station functions, we conduct some experiment on setting the value of k, we find that the experiment result is better when the value is set to 20. When the value is too small or too big, the differential of semantic vectors is unsatisfactory. Therefore, we set the value to 20 in the subsequent experiment. We have solved the problem of mobile semantics acquisition in this sub-section.

3.1.2 Location semantic mining: For obtaining the location semantics from the POI data, we conduct location semantic mining by implementing LDA model. We create a $m \times n$ matrix $\boldsymbol{M}_{\text{SPOI}}$ based on the station's POI data and standardise each row of the matrix by min–max, then we obtain the station-POI content matrix $\boldsymbol{M}^*_{\text{SPOI}} \cdot m^*_{i,j}$, and take this matrix as the input of LDA model. Finally, we obtain a $m \times k$ station-function matrix, of which each row in $m \times k$ matrix represents the k latent location semantics distribution in one station.

The function of a subway station region can be represented not only through its passenger flow mobile patterns, the static places (such as buildings etc.) are the carrier that people can engage in various activities, their categories can also reflect the function of region. Such as one region with a large proportion of schools, it will have a higher proportion to be represented as the education function. In location semantic mining of subway station regions by using the POI dataset, we use statistics to calculate the number of each POI label in a station region, so we create a $m \times t$ matrix $\boldsymbol{M}_{\text{SPOI}}$ based on the station's POI data, of which m is the number of stations, t is the number of POI category labels and the element $\boldsymbol{M}_{\text{SPOI}} \cdot m_{i,j}$ in row i column j is the number of the POI category j's label in the station i's region. In our approach, we select the POI data within 500 m of the station. However, it is not appropriate to analyse regional function of stations only using the number of POI, because there is a great difference between the absolute quantity of different POI categories. For instance, we find the quantity of 'shopping mall' POI is significantly more than other POI labels in most station region, and the 'scenic spot' POI is lesser than other POI labels. In this situation, the larger absolute quantity of 'shopping mall' POI labels than 'scenic spot' POI labels does not mean the station location semantic mining on 'scenic spot' POI labels are unimportant. In order to solve this problem, we map the numerical value of each POI category to the range of 0–1, we define this as the POI relative amount. Specifically, we standardise each row of the matrix $\boldsymbol{M}_{\text{SPOI}}$ by min − max, in the following equation:

$$\boldsymbol{M}^*_{\text{SPOI}} \cdot m^*_{i,j} = \frac{\boldsymbol{M}_{\text{SPOI}} \cdot m_{i,j} - \min(\boldsymbol{M}_{\text{SPOI}}[, j])}{\max(\boldsymbol{M}_{\text{SPOI}}[, j]) - \min(\boldsymbol{M}_{\text{SPOI}}[, j])}, \quad (1)$$

where $\min(\boldsymbol{M}_{\text{SPOI}}[, j])$ is the minimum value of the matrix $\boldsymbol{M}_{\text{SPOI}}[, j]$ column j, $\max(\boldsymbol{M}_{\text{SPOI}}[, j]$ is the maximum value of the matrix $\boldsymbol{M}_{\text{SPOI}}[, j]$ column j, of which $i = 1, 2, 3, \ldots, m$, $j = 1, 2, 3, \ldots, t$. Thus, we obtain the station-POI content matrix $\boldsymbol{M}^*_{\text{SPOI}} \cdot m^*_{i,j}$. Then, as the same with mobile semantic mining, we make an analogy to the station's location semantic as shown in Table 2.

Thus, we take the matrix $\boldsymbol{M}^*_{\text{SPOI}} \cdot m^*_{i,j}$ as the input of LDA model. Then we can obtain a $m \times k$ station-function matrix which is reflected by the static places around the station, of which m is the number of subway stations, k is the number of the latent functions. Each row of this matrix represents the distribution of k latent location semantics in one station. As the same with mobile semantics mining, we set the value of k to 20. We have solved the problem of location semantics acquisition in this sub-section.

Table 1 Analogy between station-mobile semantic mining and document-subject mining

Location-mobile semantic mining	Document-subject mining
subway station dataset	document set
station	document
function of stations	topic of documents
travel patterns within the station	words within the document

Table 2 Analogy between station-location semantic mining and document-subject mining

Station-location semantic mining	Document-subject mining
subway station dataset	document set
station region	document
function of station region	topic of documents
POIs content distribution within the station region	the distribution of a word within the document

3.2 Functional features extraction

For obtaining the final functional excavation results, we need to extract the functional characteristic of each station from the results. Before this, we define each of the mobile semantic vectors and the location semantic vectors as a characteristic of subway stations and synthesise into a matrix $\boldsymbol{M}_{SF}$. Then, we standardise this matrix by using Z-score normalisation and extract the characteristic vector of each station by using SPCA. Finally, we obtain a station-function matrix $\boldsymbol{F}$.

In previous sub-section, we have given a detailed introduction about how to obtain the station-mobile semantics reflected by passenger travel patterns and the station-location semantics reflected by the distribution of POI labels. It needs to have considered both the two semantics for knowing the distribution of the station's function because the single semantics cannot completely reflect the function of stations. For instance, we identify the station's function by the result from passenger flow data, both of the commercial region and educational region are presented the same travel pattern because of the distributions of the activity (go to work and go to school) in travel patterns are very similar. It is difficult for us to identify the function by the single station-mobile semantic, but if we take the POI labels into consideration, the accurate function will be identified easily. Consequently, the functional semantics of subway station is the combination of its mobile semantics with its location semantics.

The functional characteristic of a station is determined by the mobile semantic vector and the location semantic vector within the station. Therefore, we splice the two vectors and define each of them as a characteristic of stations, as a result, we obtain a $m \times 2k$ matrix $\boldsymbol{M}_{SF}$, of which m is the number of stations, and the mobile semantics and the location semantics both have k latent topics. The two matrices obtained after the processing of LDA model have

different average values and different dispersion degrees, because the passenger flow data and the POI data we used are heterogenous for different sources and different sizes. If we do not process each dimension in the two matrices with the same standard by conducting standardisation, the proportion of each factor will be different in the followed process of extracting features. For example, the factor with larger mean value will cover up the factor with smaller mean value, or the changes of dispersion factor will cover up the changes of small variance factor, both of them can distort the result. Therefore, we perform a Z-score normalisation on the matrix $\boldsymbol{M}_{SF}$ column. The role of Z-score normalisation is to process all the column vector into the standard normal distribution, of which its expectation $\mu = 0$, its variance $\sigma = 1$. It can dismiss the influence of data dimension on subsequent calculating, and the computing method of Z-score normalisation is given by

$$\boldsymbol{M}^*_{SF} \cdot m^*_{i,j} = \frac{\boldsymbol{M}_{SF} \cdot m_{i,j} - \mu_j}{\sigma_j}, \tag{2}$$

where μ_j is the expectation of matrix $\boldsymbol{M}_{SF}$ column j, and σ_j is the variance of matrix $\boldsymbol{M}_{SF}$ column j.

After standardisation, the dimension of the mobile semantic vector is equivalent to the dimension of the location semantic vector. At this time, we can extract the characteristic vector of each subway station by using SPCA.

At the first, principle components analysis (PCA) is a commonly used method of effective dimensionality reduction in multivariate data statistics and analysis, it only retains the component with high variance after taking linear transformation (the larger the variance, the more information the component contains), these new variables are based on the linear combination of the original characteristic variables, and they become the principle component.

Suppose there is a $n \times f$ sample set $\boldsymbol{X}$, the procedure of PCA includes: first, centralise the samples (subtract the mean), then calculate the covariance matrix $\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$, conduct the eigenvalue decomposition of this matrix $\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$, and choose the eigenvector $pc_1, pc_2, \ldots, pc_{n'}$ which corresponds to the n' maximum eigenvalues as the principle components, and at last the resultant matrix $\boldsymbol{Y} = (pc_1, pc_2, \ldots, pc_{n'})$ is obtained. Meanwhile, it defines the f vectors as the load vector of each principle component, of which these vectors are belonging to the matrix $\boldsymbol{V} = (v_1, v_2, \ldots, v_f)$ that makes the $\boldsymbol{X}\boldsymbol{V} = \boldsymbol{Y}$ true.

However, the principle components are the linear combination of all the original variables. Each original variable has non-zero load on each principle component. As a result, it is hard to distinguish and explain the true meaning of the obtained principle components in practical application.

For dealing with this situation, researchers have proposed the SPCA, which aims to make every principle component to be the linear combination of less original variables, which limits the quantity of none-zero value in the load vectors. Under this precondition, it maximises the interpretable variance by these linear combinations. As for the solution about SPCA, it can be regarded as the sparse representation of the principle component of PCA, which is a learning with a L_1 regular penalty dictionary, its purpose is shown by

$$(U^*, V^*) = \arg\min_{U,V} \frac{1}{2} \parallel X - UV \parallel_2^2 + \alpha \parallel V \parallel_1 \tag{3}$$

In the implementation of SPCA, we select the top n principal components that can be used to interpret the variance greater than 90% as the extracted final functional characteristics of subway stations, then we obtain a $m \times n'$ station-function matrix $\boldsymbol{F}$.

3.3 Subway station functional clustering

The station-function matrix $\boldsymbol{F}$ records the function vectors of each station. In order to identify which stations have the similar function, we cluster these functional vectors. Stations which have the similar function will be divided into the same cluster. In the process of selecting clustering algorithms and determining the number of clusters, there is no label cluster that can be compared. Therefore, we decide to choose an important internal indicator of clustering results which is named silhouette coefficient for evaluating clustering performance. The silhouette coefficient is calculated using the two indices below:

a is the average distance between a sample point and all other sample points in the same cluster, it reflects the cohesion degree in the cluster.

b is the average distance between a sample point and all other sample points in its nearest cluster, it reflects the separation degree between the clusters.

The formula for calculating the silhouette coefficient of a sample is given by

$$s = \frac{b - a}{\max(a, b)} \tag{4}$$

The silhouette coefficient of a sample set is the average of each sample coefficient, and the higher the value it represents, the better the effect of clustering which means the distance between the samples in the same cluster is small and the distance between clusters is large. We made multiple comparisons in the experiment for selecting clustering algorithms and the number of clusters. Finally, we decide to use k-means algorithm, cluster number is set to 10.

K-means algorithm is used to partition the minimisation squared error from the clusters after clustering, of which the number of clusters is required to set in advance. k-means algorithm also have a good performance in the large-scale dataset, and this algorithm is widely used in many different fields.

Given below is a description of k-means algorithm [25].

Given a set of observations $(X_1, X_2, \ldots, X_n)$, where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into k $(k \leq n)$ sets $S = \{X_1, X_2, \ldots, X_k\}$, which can minimise the within-cluster sum of squares (WCSS) (i.e. variance). In other words, the goal of k-means algorithm is to find the center of a given cluster, which makes each point belong to the cluster with the minimum squared deviation.

$$\arg\min_S \sum_{i=1}^{k} \sum_{x \in S_i} \parallel x - \mu_i \parallel^2 \tag{5}$$

In (5), μ_i is the mean of points in S_i. To find the optimal solution for minimising the squared error, it needs to examine all possible cluster divisions in S, which is a NP-hard problem. The approximate solution is obtained through iterative optimisation by k-means algorithm. Given an initial set of k means $m_1^{(1)}, \ldots, m_k^{(1)}$, the k-means algorithm proceeds by alternating between two steps:

Assignment: Assign each observation to the cluster whose mean generates the minimised WCSS

$$S_i^{(t)} = \{x_p : \parallel x_p - m_i^{(t)} \parallel^2 \leq \parallel x_p - m_j^{(t)} \parallel^2 \forall j, 1 \leq j \leq k\} \tag{6}$$

In (6), each x_p is assigned to one $S^{(t)}$, and even if x_p could be assigned to two or more of $S^{(t)}$.

Update: Calculate the new means to be the centroids of the observations in the new clusters

$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j \tag{7}$$

The arithmetic mean is a least-squares estimator, thus it also minimises the objective of WCSS. The algorithm has converged when the assignments no longer change. It needs to be pointed out that the solution we obtained is the locally optimal solution which

Table 3 POI labels

Label name	Description
government	all kinds of government agencies
science & education	all categories of training institutions
public utilities	kiosks, public toilets, public telephones
shopping mall	kinds of stores for shopping
scenic spot	tourist attractions, parks, squares
finance & insurance	banks, insurance companies
accommodation	various types of hotels
living services	bathing centre, barber shops, laundries
medical services	drug stores, hospitals, clinics
food services	restaurants, coffee shops
sports & leisure	entertainment venues
corporate business	companies in all industries
apartment	business offices and residential areas

Table 4 Field of the smart card transaction data in Shanghai subway

Field name	Field type	Field meaning
card ID	varchar	passenger's smart card ID
date	date	date of travel
StartTime	time	start time of the travel
StartStation	varchar	departure station
StartFee	double	inbound charges
StartX	double	departure station longitude
StartY	double	departure station latitude
EndTime	time	end time of the travel
EndStation	varchar	destination station
EndFee	double	outbound charges
EndX	double	arrival station longitude
EndY	double	arrival station latitude
distance	double	travel distance
duration	time	schedule time
interval	time	travel time

means the initial values of the centroids of clusters can influence the final result. Consequently, the algorithm often chooses different initial centroids to implement many times.

After clustering F, ten clusters are obtained, we marked them as follows $c_1, c_2, \ldots, c_{10}$, each cluster is the collection of the stations with the same function.

3.4 Subway station function classification

After obtaining the station clusters, we further add the semantic labels to each station cluster for helping people to have an intuitive understanding on its real function. However, it is important to note here that whether in traditional urban planning or in nowadays document processing, both region annotation and document annotation are very challenging questions. In fact, even now the task about the topic model concretisation is still a problem that cannot be solved very well. When we applied this process, the category of each station cannot make an analogy with any existing labels, and there is no explicit criteria for determining the function of stations. These issues definitely make this work more difficult.

In our approach, we try to analyse station function classification through the following angles:

1. *Inter-class passenger flow transfer*: It is similar to the previous calculation on the quantity of outbound and inbound passenger flow between subway stations. By analysing the characteristic of the inter-class passenger flow between different periods of time, it can help us to take the type identification. We calculate the inter-class average passenger flow per station from different periods of time, weekdays and weekends. For each category, it has four passenger flow volume-temporal distribution matrices. Specifically, in time period t, the average passenger flow from the stations in cluster c_i to the stations in cluster c_j is the total number of passenger flows from cluster c_i to cluster c_j divided by the product of total number of stations in the two clusters. We can interpret the inter-class passenger flow transfer value as the inbound and outbound passenger flow of each typical station in different period of time after abstracting the ten categories into ten typical subway stations.
2. *Geographical function proportion distribution*: Figuring out the percentage of the POI number per station within the total number of the city can enhance the accuracy of the function identification of subway stations. The proportion of the ith POI label in the geographical function of station category j is given by

$$r_{i,j} = \frac{n_{i,j}}{n_i \cdot n_j}, \tag{8}$$

where n_i is the number of category i POI, n_j is number of j category station, $n_{i,j}$ is the number of all the categories i POI in the station region of category j.
3. *Inter-cluster similarity*: We can calculate the inter-cluster cosine similarity matrix $\boldsymbol{M}_S$ based on the obtained ten clustering centre vectors $\mu_i (i = 1, 2, 3, \ldots, 10)$. Of which $\boldsymbol{M}_S$ is a 10×10 matrix, the specific calculating method of each element $\boldsymbol{M}_S \cdot m_{i,j}$ is given by

$$\boldsymbol{M}_S \cdot m_{i,j} = \cos\langle \mu_i, \mu_j \rangle \tag{9}$$

In this process of subway station function classification, the bigger the inter-cluster similarity, the more similar the function of the two clusters, and it can be taken as a valuable evaluation standard. The specific analysis and result of function classification will be introduced in the following experimental result analysis.

4 Analysis of experimental results

4.1 Data description

We conduct experiments to verify the effect of the approach we proposed by using three datasets below.

Shanghai subway line network data: We pre-process the dataset for merging the stations at the intersection of different lines. The dataset we used in this paper includes 288 stations of 14 lines, and 47 stations of them are transfer stations.

Shanghai POIs data: Each POI information of the POIs dataset includes the category of functional regions, the latitude and the longitude. In this dataset, we only use 13 POI labels (see Table 3) in this experiment, the other POI labels are ignored.

Smart card transaction data in Shanghai subway. This dataset contains the records of passenger's smart card swiping data in April 2015. The field name of each record and its interpretation are shown in Table 4.

4.2 Clustering result and function classification of subway stations

After processing the data as described in the experimental section of this paper, we get the classification result of subway stations which is shown in Fig. 3*a*. From this result, we can find the distribution of the subway stations with different functions in Shanghai has the following features:

i. Each of the important subway station is divided into a single functional category. In Fig. 3*b*, each of the two special stations is divided into the single categories, the Shanghai railway station and the people's square station are exactly located in the two most important places of Shanghai. Shanghai railway station is the important transportation junction in the city, and there are three subway lines passed through this station, thus it

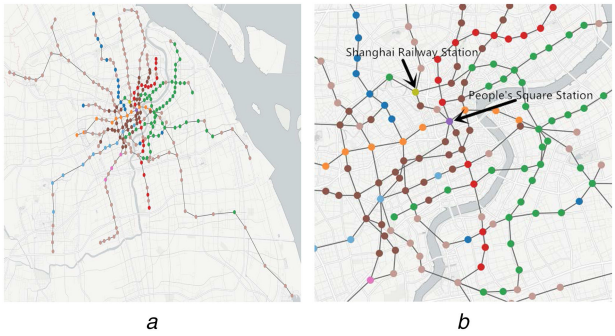


Fig. 3 *Classification results of subway stations*
(a) Global results, ***(b)*** Detailed results

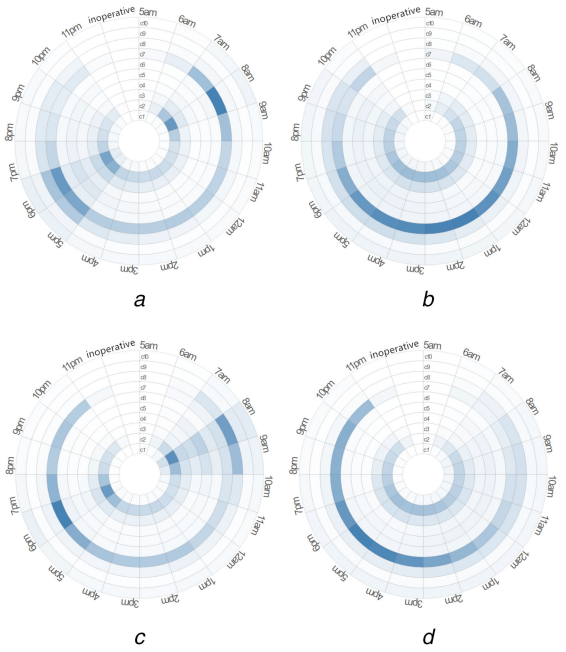


Fig. 4 *Characteristics of passenger flow transfer in tourist stations*
(a) Departure passenger flow on weekdays, ***(b)*** Departure passenger flow on weekends, ***(c)*** Arrival passenger flow on weekdays, ***(d)*** Arrival passenger flow on weekends

is also important to railway system. As for the people's square station, it is the most famous landmark in Shanghai and the centre of politics, culture, travel, economy of Shanghai.

ii. Large-scale regional division is obvious. In the visualisation above, the distribution of the light brown, puce and green colours in station classification are obvious regional, of which the puce stations are mainly in the centre region of Shanghai, the light brown stations are widely distributed in the outer suburbs of Shanghai and the green stations are stretched along the Huangpu river, most of them are located in Pudong New District.

iii. The function of the stations in the same line is similar. Despite of the accumulated stations, we find that the rest of the subway stations are located at the same line, it revealed the important line of Shanghai subway system with special functions. Such as the orange stations are located on subway line 2, and this line has connected the Hongqiao airport and the Pudong international airport, and it also has passed many bustling landmarks. We can find many significant information from the station clustering results, as for taking functional classification to these clusters, the quantified benchmarks liked inter-class passenger flow transfer, geographical function proportion distribution and the inter-cluster similarity should be taken into consideration.

Based on the clustering results above, we analyse each clusters from the three aspects following the description of the functional features extraction presented in previous section. The calculated results of the inter-category passenger flow transfer are displayed in Figs. 4–6. The number of the whole result graphs is 40. Therefore, we select three typical result graphs for displaying. It shows the geographical function proportion distribution in Figs. 7 and 8. These visualisation show the calculated results of the inter-cluster similarity matrix. In the following subway station functional classification, we will elaborate the performance and impact of each indicator in the final result.

Through the aforementioned three comprehensive considerations, we have made the following nine classifications of the subway station's function. It should be noticed that we only take the most obvious or distinctive function of a station classification as its identifier, but this does not mean the classification has only one function. In reality the function of a station is diverse and fuzzy.

4.2.1 People's square station category (c_7)*:* The people's square is the political and economic centre of Shanghai. People's square station has become a special and single station classification. From the angle of inter-class passenger flow transfer, there are enormous passenger flow volume in all kinds of subway station especially in holidays, so it can be seen that this station is a tourist destination. From the angle of geographical function proportion, the total number of construction facilities nearby the people's square are far away from the any other subway stations, so it can be seen that the region around the people's square is the most dense region of Shanghai's construction facilities, and with a full range of functions.

4.2.2 Transportation junction category (c_3, c_8)*:* We incorporate c_3 and c_8 into this functional classification, because both of them assumed the similar function as transportation junction. The c_3 is the classification which represents Shanghai railway station, and c_8 is the classification which includes Shanghai south railway station and the two nearby stations after clustering. The passenger flow values of these two clusters are pretty large, but from the angle of geographical functions, the number of construction facilities around the stations in c_8 is more less than Shanghai railway station. Therefore, despite these two clusters are divided into one important function classification, but c_8 has the smaller ability on providing the multiple functions and services than Shanghai railway station c_7. By the way, the railway station becomes the typical and single classification in our result but not the airport, this result also revealed that although Shanghai is a first-tier city, the trains are the major transportation chosen for most people in Shanghai.

4.2.3 Tourist recreation category (c_2)*:* The main distribution of the stations in this classification is the tourist recreation regions in the centre of Shanghai. From the angle of the inclusion stations, regions which attract tourists are all located in cluster (c_2), such as Hongqiao airport, East Nanjing Road, Jing'an Temple and so on. From the angle of inter-class passenger flow transfer, stations of tourist recreation are appeared afternoon peak and evening peak, especially on weekends. From the view of inter-cluster passenger flow transfer, the transfer peak of the stations in c_2 is reached in the afternoon and evening, especially on weekends. From the view of the geographical function proportion distribution, c_2 takes a large proportion in scenic spots. The similarity between the tourist recreation categories and the People's Square is the highest than others, so we can regard the People's Square as the most typical tourist recreation station.

4.2.4 Economical culture category (c_{10})*:* The economical culture stations are located in the most developed areas of Shanghai science and education culture and economic output, which includes Fudan University, Shanghai Jiaotong University, Tongji University, Shanghai Library and so on. There is a typical passenger flow of these subway stations that show the feature of go out early and return till late. The proportion of science and education POI is only a little lower than the People's Square Station, but it has a large number of stations (42 stations), such that it also has the large total

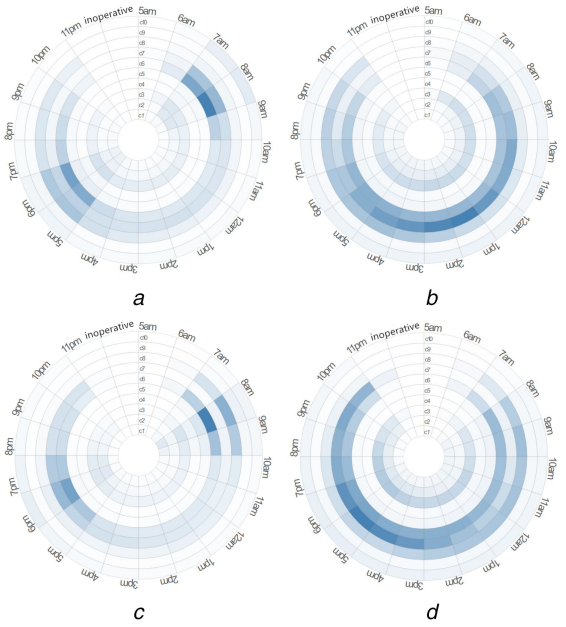


Fig. 5 *Characteristics of passenger flow transfer in commercial corporation stations*

(a) Departure passenger flow on weekdays, ***(b)*** Departure passenger flow on weekends, ***(c)*** Arrival passenger flow on weekdays, ***(d)*** Arrival passenger flow on weekends

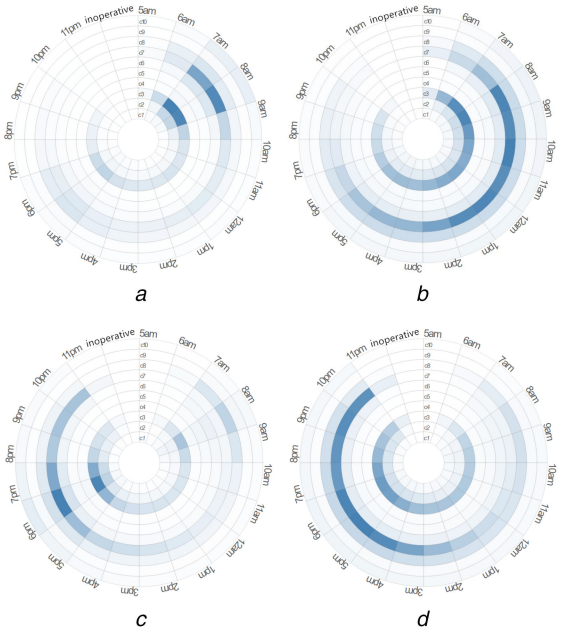


Fig. 6 *Characteristics of passenger flow transfer in ordinary residential stations*

(a) Departure passenger flow on weekdays, ***(b)*** Departure passenger flow on weekends, ***(c)*** Arrival passenger flow on weekdays, ***(d)*** Arrival passenger flow on weekends

number of science and education culture POIs. Besides the education, category c_{10} regions are the economical centre of Shanghai, which the financial services and business companies account for a high proportion. This category also has the high similarity with the categories of tourist recreation and transportation junction, it is the mainstay of the economic and cultural development in Shanghai.

4.2.5 Commercial company category (c_6)*:* Subway stations of the commercial company category are within the city and suburban districts. Time distribution of passenger flow of these stations is pretty dispersed on rest days but it reached the peak in the morning and evening on weekdays. We also find that there are many intra-

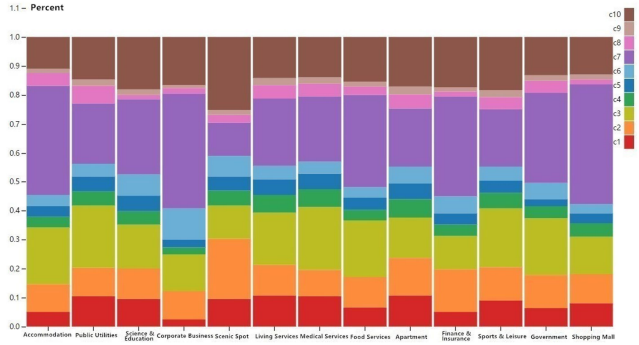


Fig. 7 *Geographical function proportion distribution*

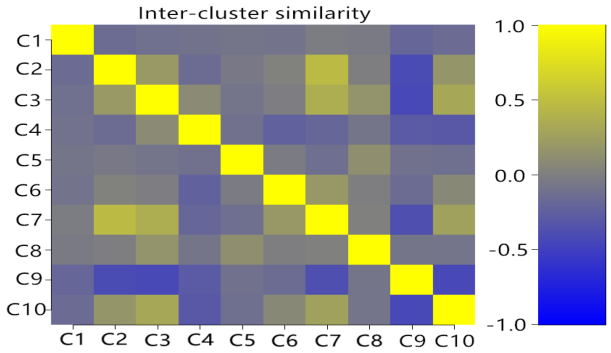


Fig. 8 *Visualisation of the inter-cluster similarity matrix*

category passenger flow shifts in the morning (see Fig. 5), which represent the passengers travelled between the stations of c_6, it also reflects the characteristic of mutual communication between commercial companies. Incorporate business has the highest POI proportion within this category, because there is some overlap in functionality, which is similar to c_{10}.

4.2.6 Developed residential area category (c_1)*:* Subway stations of this category belonged to high-grade residential areas, which are distributed in developed regions of Shanghai such as Hongkou, Jing'an district, Huangpu district, Minhang district and so on. In this category, it appears a distinct passenger flow peak of the morning out and the evening back on weekdays, but on rest days, passenger flow peak is more dispersive. Commercial apartments, public facilities and living services occupy the larger proportion in this region. Conversely, commercial companies have the least proportion.

4.2.7 Normal residential area category (c_9)*:* Cluster c_9 has the most quantity in all clusters, it is widely distributed in the nearby suburban districts of Shanghai, which is relatively low-end residential areas comparing with category c_1. It is similar to c_1 that has typical travel pattern of residential areas (see Fig. 6). However on the POI proportion, there are fewer POIs on average near each subway station. We summarise two reasons resulting in this consequence: one is the density of suburban buildings and facilities are far below the centre areas of Shanghai; the other is the sampling of the POI dataset we used mainly focused on centre areas of Shanghai. So that the POIs records of surrounding areas are in deficiency states, many information is not been recorded. On the aspect of inter-cluster similarity, we can find that the stations in normal residential area are opposite to the stations in clusters c_2 and c_{10}. Although the regions from these two clusters represented the centre of economical culture and tourist recreation, it has a negative similarity between cluster c_9 and the two clusters c_2 and c_{10}.

4.2.8 New developing district category (c_4)*:* Most of subway stations in cluster c_4 are located in Pudong new area of eastern Huangpu River, but it does not conclude the most developed Lujiazui district (in tourist recreation category) of Pudong new

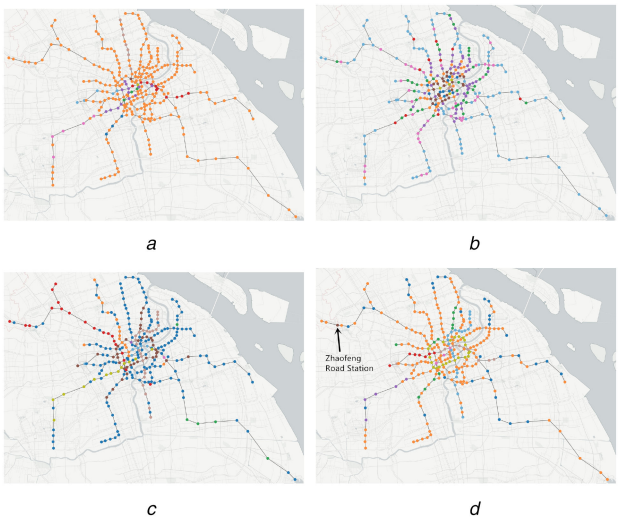


Fig. 9 *Characteristics of passenger flow transfer in ordinary residential stations*
(a) Subway station identification only using mobile semantics, ***(b)*** Subway station identification only using location semantics, ***(c)*** Clustering result by using TF-IDF to process POI dataset, ***(d)*** Clustering result without conducting SPCA

area. This feature illustrates that the cluster c_4 is not represented as the developed subway station regions but represented as the developing station regions, and it is also becomes a valid proof of the rationality of the results of this method. As a developing new district, the proportion of all kinds of POI of station region in cluster c_4 is not high, but it is still better than cluster c_9, which shows the potential of a new district.

4.2.9 Mixed category (c_5): Cluster c_5 is mainly distributed in Jiading district of Shanghai. There is no shortage of geographical functions, but they are also not conspicuous. If the station's classification is sorted from the modernisation to the backward sort, cluster c_5 is located in the median interval. It is a typical mixed station category which has high similarity with category c_9. Actually, subway stations in Shanghai suburbs generally presented this feature. Subway stations in Jiading district turn into a category alone that is relative to algorithmic initialisation randomness of k-means. Therefore, cluster c_5 only existed in Jiading district in this clustering result, but actually it is the representative of all the suburban and functional fuzzy stations.

5 Performance comparison

In this section, we mainly show and analyse the function recognition results of the contrast experiment. Specially, there is no standard answer to the questions of station functional identification, nor is there a right or wrong judgement, when we analyse the result of classification, we mainly focus on whether the classification of some typical subway stations is to be ideal and take this as the basis of the evaluation.

In regard to the design of this experiment, the method proposed in this paper is not only a single algorithm, but also a set of urban computing processes. Therefore, we design along the lines of a change in a single variable on comparing with other algorithms under the same framework of FISS. We replace a step of our method into another algorithm or just remove it, the rest steps is the same, and we compare the relative merits of results from the original method and the modified methods. We display the contrast experiment in the following subsections below.

5.1 Mining single semantics

In order to show the validity of the united mobile semantics and location semantics mining, we conduct the experiment that only mining single mobile semantics or location semantics (see Figs. 9*a* and *b*).

It is observed that many stations are clustered to the same category (the orange stations) with the similar travel pattern that only mining mobile semantics. They cannot show the otherness of building facilities between each station's region. Then, we only analyse the location semantics, we can find that the distribution of station clusters is blended, and it is particularly obvious in urban centre regions, the regional distribution characteristics are not obvious so that the results can be interpreted to a low degree.

5.2 Using TF-IDF to process POI dataset

TF-IDF is a general term weighting scheme for extracting keywords in documents [26]. It can be used to calculate the key POI category of a station in POI dataset, as for the $m \times n$ station-POIs matrix $\boldsymbol{M}_{\text{SPOI}}$ in (1). The formula of the TD-IDF value $v_{i,j}$ of the element $\boldsymbol{M}_{\text{SPOI}} \cdot m_{i,j}$ is given by

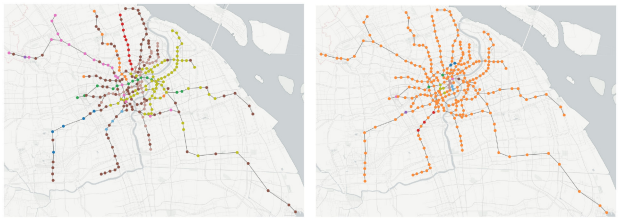
$$v_{i,j} = \frac{\boldsymbol{M}_{\text{SPOI}} \cdot m_{i,j}}{N_i} \times \log \frac{m}{\| \{i | \boldsymbol{M}_{\text{SPOI}} \cdot m_{i,j} \neq 0\} \|} \tag{10}$$

In this equation, $N_i = \Sigma_{k=1}^{t} \boldsymbol{M}_{\text{SPOI}} \cdot m_{i,k}$. $\boldsymbol{M}_{\text{SPOI}} \cdot m_{i,j}/N_i$ indicates the frequency (term frequency, TF) at which the POI category j appears in the station region i, $\log(m/\| \{i | \boldsymbol{M}_{\text{SPOI}} \cdot m_{i,j} \neq 0\} \|)$ is the inverse document frequency (IDF) of one POI category, which indicates the rarity of this POI category, and TF-IDF measures the importance of the POI category in a station region by multiplying these two items. We replace the POI content distribution to the TF-IDF results, and take it as the input of the LDA model, other steps remain unchanged. The clustering result is shown in Fig. 9*c*.

We can find that the important categories (such as economical culture, tourist recreation etc.) are not obvious in this result. The overall effect and interpretability is inferior to the result in this paper. The result by using TF-IDF to process POI dataset is not ideal in this experiment. The problem we found is in IDF part. As the POIs dataset we used is dense, the absolute numbers of partial POI categories are not high, and their distributions are rarely to be zero, therefore, is makes less obvious difference in the TF-IDF value of the whole POI categories. At this time, TF-IDF is only worked as TF, the frequency of the POI categories in the same station area cannot well reflect the importance of single POI category. This is also the reason why the POI content distribution used in this paper is better.

5.3 Experiment without conducting SPCA

The clustering result is shown in Fig. 9*d*. This result is apparently not ideal. Not only the special subway stations (such as Shanghai Railway Station, People's Square Station etc.) are not highlighted, but also most of the subway stations in the downtown and in the city are clustered into the same category (the orange spots) which cannot be distinguished. We have marked the Zhaofeng Road Station with an arrow. This station becomes a category because of



a b

Fig. 10 *Subway stations clustering by using AP algorithm and MeanShift algorithm*
(a) Clustering result by using AP algorithm, ***(b)*** Clustering result by using MeanShift algorithm

the remaining noise data without conducting SPCA. We have introduced the necessity and the advantage of conducting SPCA in Section 3. The rationality and the validity of SPCA can be proved through this contrast method.

5.4 Subway stations clustering by using AP algorithm and MeanShift algorithm

In data mining and statistics, AP is a clustering algorithm based on the concept of 'message passing' between data points, the 'real-valued messages' are exchanged between data points until a representative set of exemplars and their corresponding clusters gradually emerge [27]. The information transmitted by data points represents the appropriateness that an exemplar becomes the representative exemplar among the others, and it is updated by obtaining values from other data points, thus, AP can find the most representative exemplars by keeping updating and iterating until they are converge. The number of clusters in AP is unnecessary to be assigned but based on the dataset used.

The description of AP is given below [28].

Let x_1 through x_n be a set of data points, and let s be a function that quantifies the similarity between any two points, such that $s(x_i, x_j) > s(x_i, x_k)$, where x_i is always more similar to x_j than to x_k.

There are two types of information passed by two data points: $r(i, k)$ is the values of the 'responsibility' matrix $\boldsymbol{R}$, which represents the attraction information that x_k is better to serve as the exemplar for x_i than other exemplars for x_i. $a(i, k)$ are contained in the 'availability' matrix $\boldsymbol{A}$, which represents the associated information, reflecting the appropriateness that x_i to choose x_k as its exemplar, it also has considered that the other points could also choose x_k as an exemplar. The data point will be chosen as the exemplar of others point if it satisfies the following two conditions: it is similar enough with many data points. It is chosen as the exemplar of many data points.

While we make the station clustering by using AP algorithm, as shown in Fig. 10*a*, the special subway stations (People's Square Station, Shanghai Railway Station etc.) are not separated into single category. Moreover, the stations in centre areas and suburb are also not separated. Both of them are large deviation with the actual situation. The clustering result by using MeanShift algorithm is shown in Fig. 10*b*. Only a few stations each become the single category. The other stations (the orange spots) are clustered to the same category. The clusters which include the few stations have no universality, and the cluster with the large amount stations is lacked of specificity. It is obvious that these two algorithms are not suitable for the situation in this paper. Actually, after testing many different algorithms in the experiment, we choose the *k*-means algorithm because of its flexible efficiency, the best performance and the highest silhouette coefficient.

From the comparison methods above, it is observed that the approach in this paper can produce convincing functional identification results. It is more efficient in reality about regional function identification of subway station.

6 Conclusion

Subway stations are often located in the regions where the passenger flow and the geographical functions are dense. It can provide quite a few references for the construction planning in urban transportation and regions by identifying the function the subway station carried. Meanwhile, with the rapid growth of urban city data (such as passenger flow data, geographical data etc.), all of them contain a wealth of information. Thus, we can obtain the functional semantics of subway stations by using data mining.

We have proposed an approach FISS for mining the functions of subway stations under the urban computing framework by using the subway passenger travel information and the station regional POI information, and the validity of this approach has been proved by conducting the experiment.

In future work in this line, we will address the following issues:

i. We find that the function of the stations in the same subway station line appears very similar. In our further work, we will focus on revealing the deeper correlation between the function of subway stations and the function of subway lines.
ii. The categories of POI dataset we used are not detailed enough, parts of data are incomplete. The veracity of function mining can be further improved by using the full land utilisation data.

7 Acknowledgments

This work was partially supported by the Natural Science Foundation of Liaoning Province, China under grant no. 201602154, and the Dalian Science and Technology Planning Project under grant nos. 2015A11GX015 and 2015R054.

8 References

[1] Ning, Z., Xia, F., Ullah, N., *et al.*: 'Vehicular social networks: enabling smart mobility', *IEEE Commun. Mag.*, 2017, **55**, (5), pp. 16–55
[2] Xia, F., Wang, J., Kong, X., *et al.*: 'Exploring human mobility patterns in urban scenarios: A trajectory data perspective', *IEEE Commun. Mag.*, 2017
[3] Kong, X., Xu, Z., Shen, G., *et al.*: 'Urban traffic congestion estimation and prediction based on floating car trajectory data', *Future Gener. Comput. Syst.*, 2016, **61**, pp. 97–107
[4] Kong, X., Song, X., Xia, F., *et al.*: 'Lotad: long-term traffic anomaly detection based on crowdsourced bus trajectory data'. World Wide Web, 2017, pp. 1–23
[5] Kong, X., Xia, F., Wang, J., *et al.*: 'Time-location-relationship combined service recommendation based on taxi trajectory data', *IEEE Trans. Ind. Inf.*, 2017, **13**, (3), pp. 1202–1212
[6] Yuan, N.J., Zheng, Y., Xie, X., *et al.*: 'Discovering urban functional zones using latent activity trajectories', *IEEE Trans. Knowl. Data Eng.*, 2015, **27**, (3), pp. 712–725
[7] Karlsson, C.: 'Clusters, functional regions and cluster policies'. JIBS and CESIS Electronic Working Paper Series (84), 2007, pp. 1010–1018
[8] Unsalan, C.: 'Measuring land development in urban regions using graph theoretical and conditional statistical features', *IEEE Trans. Geosci. Remote Sens.*, 2007, **45**, (12), pp. 3989–3999
[9] Zhi, Y., Li, H., Wang, D., *et al.*: 'Latent spatio-temporal activity structures: a new approach to inferring intra-urban functional regions via social media check-in data', *Geo-spat. Inf. Sci.*, 2016, **19**, (2), pp. 94–105
[10] Assem, H., Xu, L., Buda, T.S., *et al.*: 'Spatio-temporal clustering approach for detecting functional regions in cities'. 2016 IEEE 28th Int. Conf. on Tools with Artificial Intelligence (ICTAI), 2016, pp. 370–377
[11] Kraft, S., Marada, M.: 'Delimitation of functional transport regions: understanding the transport flows patterns at the micro-regional level', *Geografiska Ann. B, Hum. Geography*, 2017, **99**, (1), pp. 79–93
[12] Fan, J., Chen, T., Lu, S.: 'Unsupervised feature learning for land-use scene recognition', *IEEE Trans. Geosci. Remote Sens.*, 2017, **55**, (4), pp. 2250–2261
[13] Yin, J., Soliman, A., Yin, D., *et al.*: 'Depicting urban boundaries from a mobility network of spatial interactions: a case study of Great Britain with geo-located twitter data', *Int. J. Geogr. Inf. Sci.*, 2017, **31**, (7), pp. 1293–1313
[14] Rudinac, S., Zahálka, J., Worring, M.: 'Discovering geographic regions in the city using social multimedia and open data'. Int. Conf. on Multimedia Modeling, 2017, pp. 148–159
[15] Qi, G., Li, X., Li, S., *et al.*: 'Measuring social functions of city regions from large-scale taxi behaviors'. 2011 IEEE Int. Conf. on Pervasive Computing and Communications Workshops (PERCOM Workshops), 2011, pp. 384–388

[16] Sarkar, S., Chawla, S., Parambath, S.P., *et al.*: 'Effective urban structure inference from traffic flow dynamics', *IEEE Trans. Big Data*, 2017, **3**, (2), pp. 181–193
[17] Bhaskar, A., Chung, E.: 'Passenger segmentation using smart card data', *IEEE Trans. Intell. Transp. Syst.*, 2015, **16**, (3), pp. 1537–1548
[18] Zhao, J., Zhang, F., Tu, L., *et al.*: 'Estimation of passenger route choice pattern using smart card data for complex metro systems', *IEEE Trans. Intell. Transp. Syst.*, 2017, **18**, (4), pp. 790–801
[19] Zhang, F., Zhao, J., Tian, C., *et al.*: 'Spatiotemporal segmentation of metro trips using smart card data', *IEEE Trans. Veh. Technol.*, 2016, **65**, (3), pp. 1137–1149
[20] Itoh, M., Yokoyama, D., Toyoda, M., *et al.*: 'Visual exploration of changes in passenger flows and tweets on mega-city metro network', *IEEE Trans. Big Data*, 2016, **2**, (1), pp. 85–99
[21] Ni, M., He, Q., Gao, J.: 'Forecasting the subway passenger flow under event occurrences with social media', *IEEE Trans. Intell. Transp. Syst.*, 2017, **18**, (6), pp. 1623–1632
[22] Blei, D.M., Ng, A.Y., Jordan, M.I.: 'Latent dirichlet allocation', *J. Mach. Learn. Res.*, 2003, **3**, (Jan), pp. 993–1022
[23] Girolami, M., Kabán, A.: 'On an equivalence between plsi and lda'. Proc. of the 26th annual international ACM SIGIR Conf. on Research and Development in Informaion Retrieval, 2003, pp. 433–434
[24] 'Latent dirichlet allocation'. Available at https://en.wikipedia.org/w/index.php/, accessed 22 August 2017
[25] 'K-means clustering'. Available at https://en.wikipedia.org/w/index.php/, accessed 20 August 2017
[26] Paik, J.H.: 'A novel tf-idf weighting scheme for effective ranking'. Proc. of the 36th international ACM SIGIR Conf. on Research and Development in Information Retrieval, 2013, pp. 343–352
[27] Frey, B.J., Dueck, D.: 'Clustering by passing messages between data points', *Science*, 2007, **315**, (5814), pp. 972–976
[28] 'Affinity propagation'. Available at https://en.wikipedia.org/w/index.php/, accessed 21 August 2017